Nucleic Acids Research, 2008, Vol. 36, Database issue ***D641–D645***
doi:10.1093/nar/gkm873

The cell cycle DB: a systems biology approach to cell cycle analysis

Roberta Alfieri[1,2,*], Ivan Merelli[1], Ettore Mosca[1] and Luciano Milanesi[1]

[1]Istituto di Tecnologie Biomediche—Consiglio Nazionale delle Ricerche, via F.lli Cervi 93, Segrate, Milano, Italy and [2]CILEA, Consorzio Intrauniversitario per l'Elaborazione Automatica, via R. Sanzio 4, Segrate, Milano, Italy

Received August 2, 2007; Revised September 30, 2007; Accepted October 1, 2007

ABSTRACT

The cell cycle database is a biological resource that collects the most relevant information related to genes and proteins involved in human and yeast cell cycle processes. The database, which is accessible at the web site http://www.itb.cnr.it/cellcycle, has been developed in a systems biology context, since it also stores the cell cycle mathematical models published in the recent years, with the possibility to simulate them directly. The aim of our resource is to give an exhaustive view of the cell cycle process starting from its building-blocks, genes and proteins, toward the pathway they create, represented by the models.

INTRODUCTION

The cell cycle is a crucial event in biology that consists in a series of repeated events allowing the cell to grow and duplicate correctly. The study of the cell cycle involves the knowledge of a large number of genes and networks of protein interactions: thus a typical systems biology approach can be applied to study this process in order to verify the impact that differently regulated genes can have in normal cells and in cancer cells.

The key elements of systems biology studies are the models, which can be defined as abstract representations of biological components and processes in order to mathematically describe their structural and dynamical properties. The mathematical modelling of a biological process allows a systemic description that helps to highlight some features such as the emergent properties that could be hidden when the analysis is performed only from a reductionist point of view.

Moreover, in modelling complex systems, a complete annotation of all the components is equally important to understand the interaction mechanism inside the network. For this reason the integration of data regarding the different components of each model has high relevance in systems biology studies.

In this biological context we developed the cell cycle database, a data integration system that collects information about genes, proteins and models of different organisms' cell cycle network (Figure 1). We primary considered cell cycle information from humans since we intend to create a resource to support biomedical studies in the context of cancer research. Then we extended the database content toward the budding yeast cell cycle because of the large number of models available for this organism. According to this choice, the data integration concerns all genes and proteins involved in the cell cycle models of both the budding yeast *Saccharomyces cerevisiae* and the *Homo sapiens*. This information is taken from the most recent literature and plays a crucial role to contextualize behaviour of each cell cycle component.

DATABASE CONTENT

The database is structured in two closely related sets of tables that have been populated using different strategies. The first part related to genes and proteins annotation has been populated using a pipeline for the collection of relevant information concerning the components involved in mammalian and budding yeast cell cycle processes. The second part related to models has been populated through a system developed for the storage and the simulation of cell cycle mathematical models developed in a systems biology context.

Detailed information about gene, protein and model data stored in the cell cycle database is presented in Table 1.

Source data for genes and proteins

We started integrating information relying on data collected from KEGG (1) and Reactome (2). Indeed, these resources, even if not specific to the cell cycle, represent an important starting point for information about genes and proteins of the budding yeast *S. cerevisiae* and the *H. sapiens*. These species were chosen since they display evolutionary correlation in the regulatory mechanism such as the cell cycle (3).

*To whom correspondence should be addressed. Tel: +22 642 2600; Fax: +22 642 2600; Email: roberta.alfieri@itb.cnr.it

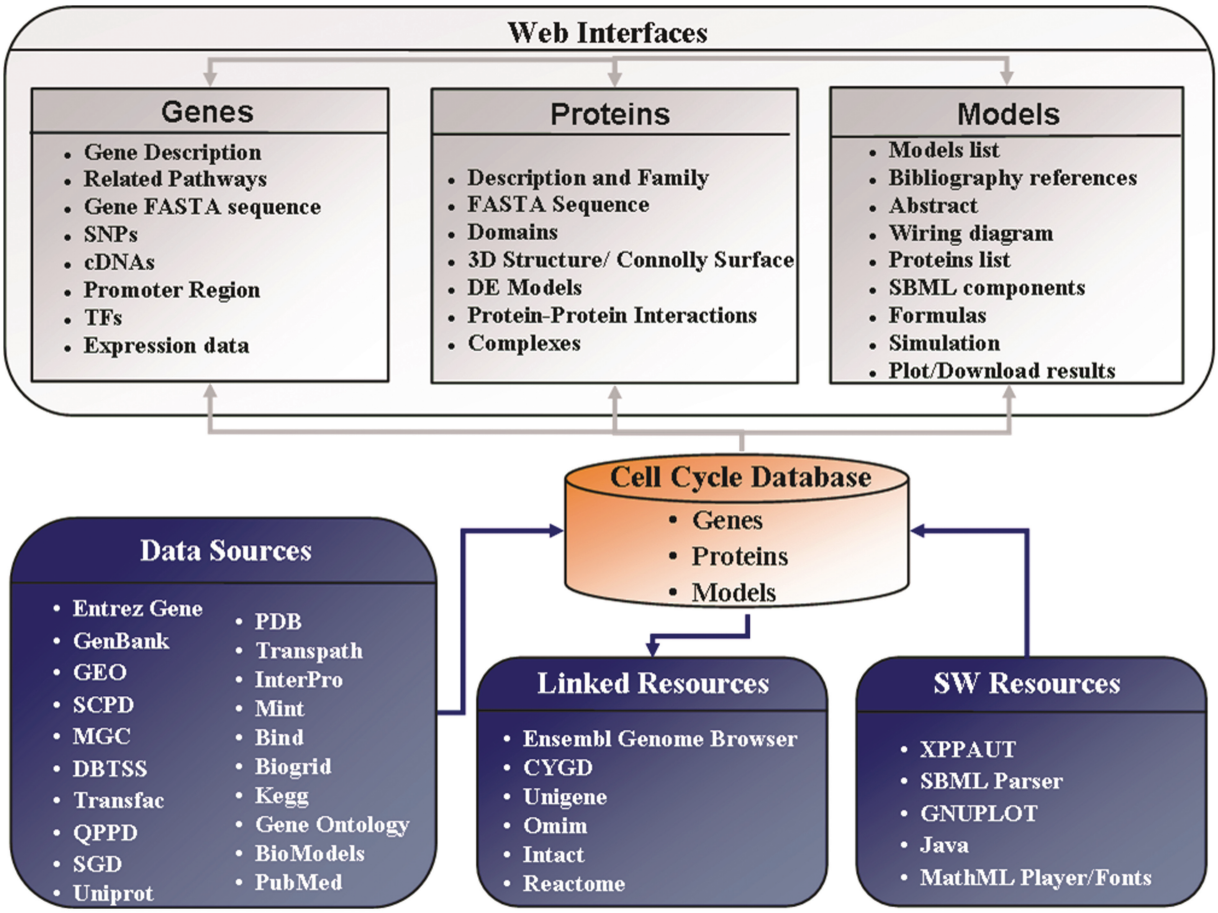

Figure 1. Architecture of the Cell Cycle Database. The core of the resource is a data warehouse system which collects data from the most relevant bioinformatics resources (Data Sources) and provides links to external resources (Linked Resources). The Cell Cycle Database comprises user-friendly interfaces which show information and, in general, provide the functions of the resource, even relying on third party free software (SW).

Table 1. Statistical information about the Cell Cycle Database content

Number of genes	559
H. sapiens	456
S. cerevisiae	103
Number of proteins	559
H. sapiens	456
S. cerevisiae	103
Number of models	26
Mammalian and Xenopus laevis	12
Yeast	14

Note: In this table the total number of genes, proteins and models listed for each organism is provided. The database stores models from 'Yeast' (i.e. *S.cerevisiae* and *S.pombe*), from mammalian and *X.laevis*.

The database contains human and yeast genes involved in the complete cell cycle pathway (cell growth pathway) and in the MAP kinase signalling pathway (a signal transduction pathway strictly related to the cell cycle). Moreover, the cell cycle database contains the human genes involved in the apoptosis pathway (cell death pathway) taken from KEGG, and it also integrates more specific information related to mitotic and checkpoint pathways from Reactome.

Cell cycle models primary data

The models list has been assembled searching literature and browsing many specific online resources. All the models relevant to cell cycle studies have been collected in the database using an XML file encoded with the Systems Biology Markup Language (SBML) (4). In particular, a number of models, for which the SBML file is available in BIOMODELS (5) or from authors websites, have been directly integrated in the cell cycle database. Published models not yet implemented in SBML have been manually encoded in SBML using JigCell Model Builder (6). All the SBML files stored in cell cycle database have been validated through the Systems Biology Workbench (SBW) SBML validator.

Up to now our resource contains 26 models, as reported in Table 1; among them, 13 have the related SBML file and for 12 of them is possible to run simulations.

Database implementation

The cell cycle database has been implemented using a relational database managed by a MySQL server. A 'data warehousing' approach has been chosen to develop the resource using a snowflake schema (7) to

organize the data. The 'data warehousing' approach is used to collect different types of data from external resources in a unique database system: in this way all data have the same format making the query system easier and faster.

The cell cycle database system consists in a series of programs written in Perl used to retrieve the data from several different external databases, to transform and load them into the warehouse data model. This process is possible using a 'snowflake' schema, a method of storing data in a relational database which presents a 'core table', where main data about yeast and human genes are stored. The 'core table' is connected to many 'external tables', where auxiliary data about genes, proteins and models are stored. This schema is particularly useful for database updating: when a new entry is inserted in the core table, all the external tables will be updated 'in cascade', while when a new entry is inserted in one of the external table no inward updating occurs.

Other resources are essentially linked to our database through public IDs, in order to gain further information and to make the integration as complete as possible. The list of the integrated and linked resources is shown in Figure 1.

APPLICATION AND WEB INTERFACE

The cell cycle database is accessible at (http://www.itb.cnr.it/cellcycle). The web interface is made up of a set of HTML pages dynamically generated from PHP scripts, in order to retrieve information about genes and proteins related to the cell cycle process in a specific report created for each gene and protein. Moreover, a search related to the cell cycle models stored in the database is possible: users can retrieve the list of the mathematical models, choose one of them and visualize the related information on the web pages.

Gene and protein report

Users browsing the website searching for annotations of the components of the cell cycle are redirected to specific reports.

Besides the common query possibilities (by gene/protein name, by keywords and by IDs of most common biological resources), users can query the database using the BLAST algorithm (8), which is particularly useful in order to discover similarities among unknown cell cycle putative genes and the database content.

The gene report lists all the information related to each gene that is stored in the database, starting from the basic gene description, its sequence and its corresponding protein, but it also includes more specific information, such as the list of the SNP characterizing that gene, or the list of cDNA and isoform. Furthermore, in the gene report, particular attention is given to the information related to the promoter regions and to the transcription factors specific for each yeast and human gene, in order to facilitate research on cell cycle gene regulation. We also provide links to experimental data on gene expression taken from the gene expression omnibus (GEO) repository in order to present as much supplementary information as possible concerning the cell cycle genes. Since the regulation of cyclin-dependent kinases (CDK) characterizes the most crucial events of the cell cycle (9), we supply additional information about kinase genes by using the link to the KinWeb database (10).

As far as the protein report is concerned, particular attention is given to the network of protein–protein interactions involved in the cell cycle. The database contains protein–protein interactions taken from several resources making the information on the cell cycle interaction network as complete as possible. In the protein report, the graphical visualization of the domains from the InterPro database is provided. Users can also directly visualize the protein structure and the related Connolly surface (11) according to PDB data, using the Java 3D applet. Moreover, for each protein we provide information on the models in which it is involved, a list of the published models is available directly in the protein report with a direct link to the specific model report discussed in the following section.

More information both for genomics and proteomics integrated and linked resources is available in the help pages of the cell cycle DB website.

Model report

Using the web interface it is possible to retrieve model-related information and a pipeline has been implemented to deal with the mathematical part of the models, in order to solve the ordinary differential equations systems that describe the biological processes. Using this system it is possible both to visualize the mathematical description of the model and to run simulations varying initial conditions of state variables and parameters.

Each model is presented in a report structured in three sections: the publication data, the SBML data structure and the numerical simulation part. The first section contains the detailed publication data, the diagram of the model, the related XML file and the list of all the proteins involved in the model that are linked to the related cell cycle database protein report.

In the SBML data structure section, users can explore the SBML components of the selected model including its mathematical expressions. Mathematical formulas within the SBML models are expressed using the Mathematical Markup Language (MathML or MML) (12). To view the expressions on the web, our resource relies on a XHTML + MathML page. This technology allows the generation of high-quality documents in which mathematical expressions are treated as text in a HTML web page: it is possible to use the browser functions to find strings inside mathematical expressions and to change their size, operations that are usually not possible using images to represents formulas (Figure 2a).

The simulation section allows users to simulate a model using the software XPPAUT (13) and to plot results on the fly in order to capture the dynamical properties of the biological process. In order to enable the simulation, this section lists the model species (state variables, typically protein concentrations), its parameters (such as kinetic

A

$$\frac{d}{dt}D(t) = \text{"aD"}\frac{\text{"GF"}}{\text{"k1"}^{-1}+\text{"GF"}} - \text{"dD"}\,E\,D$$

$$\frac{d}{dt}E(t) = \text{"aE"}\left(\frac{\text{"GF"}}{\text{"k2"}^{-1}+\text{"GF"}} + \text{"aF""theta"}\right) - \text{"dE"}\,X\,E$$

$$\frac{d}{dt}R(t) = \frac{\text{"pX"}(\text{"RT"}-\text{"RS"}-R)X}{\text{"qX"}+\text{"RT"}-\text{"RS"}-R+X} - \text{"Mass\_Action\_2"}(\text{"pS"}, R, \text{"theta"})$$

$$\frac{d}{dt}\text{"RS"}(t) = \text{"Mass\_Action\_2"}(\text{"pS"}, R, \text{"theta"}) - \frac{\text{"pD"}\,\text{"RS"}\,D}{\text{"qD"}+\text{"RS"}+D} - \frac{\text{"pE"}\,\text{"RS"}\,E}{\text{"qE"}+\text{"RS"}+E}$$

$$\frac{d}{dt}\text{"theta"}(t) = \frac{\text{"pD"}\,\text{"RS"}\,D}{\text{"qD"}+\text{"RS"}+D} + \frac{\text{"pE"}\,\text{"RS"}\,E}{\text{"qE"}+\text{"RS"}+E} + \text{"atheta"}\left(\frac{\text{"GF"}}{\text{"k3"}^{-1}+\text{"GF"}} + \text{"fe""theta"}\right) - \text{"Mass\_Action\_2"}(\text{"pS"}, R, \text{"theta"}) - \text{"dtheta"}\frac{X}{\text{"qtheta"}+X}\,\text{"theta"}$$

$$\frac{d}{dt}X(t) = \text{"aX"}\,E + f\,\text{"theta"} + gX^2E - \text{"dX"}\,X$$

B

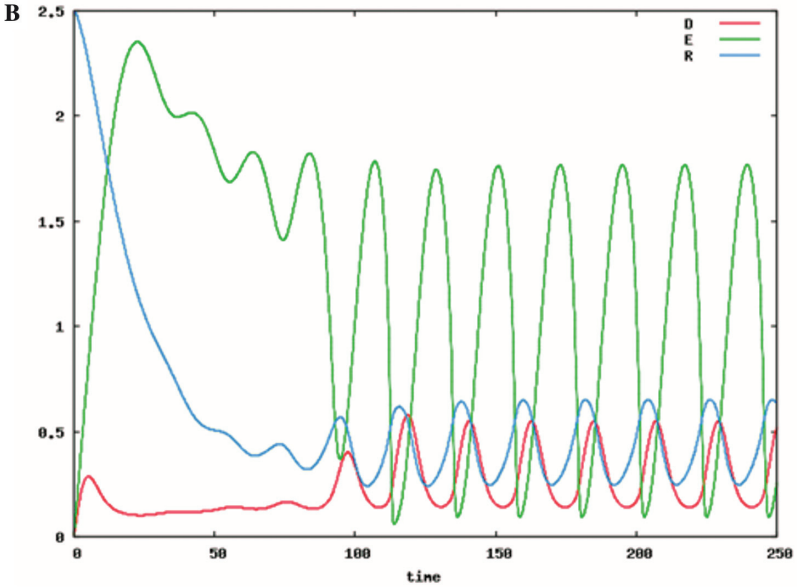

Figure 2. Model of Bai et al 2003 [15]. Examples of two objects shown by the user friendly interfaces in the ''models section'' of the Cell Cycle Database: (**A**) the ordinary differential equations as they appear on the web interface; (**B**) plot of time courses of three state variables of the model.

constants), its algebraic rules and XPPAUT internal options, using default values. Users can change the initial values in order to analyse the different natures of the system dynamics, performing sensitivity and bifurcations analysis. Once the computation is completed, users can download XPPAUT input and output files and plot results, as shown in Figure 2b. The web interface allows users to plot both time courses and phase diagrams for each model after the simulation step. Results are shown with images exported by GNUPLOT (14), the popular portable command-line function plotting software.

CONCLUSION AND FUTURE DEVELOPMENT

The cell cycle database is a freely available resource developed with the aim to support systems biology research on the cell cycle. The database aims to become a useful resource for collecting all the information related to actual and future models of this network.

The added value of our work consists in the annotation of mathematical models, which is achieved through the integration of both the model objects (species) and protein reports, where the most relevant data for each protein are presented in a standard format.

In this way the resource is useful both for retrieving information about cell cycle model components and for analyzing their dynamical properties. The cell cycle database can be used to find system-level properties, such as stable steady states and oscillations, by coupling structure and dynamical information about models.

Future trends

We plan to improve the dataset toward the analysis of how genes and proteins interact with each other to highlight the transcriptional activation and the feedback loops in the cell cycle network. We will also include the cell cycle information of other higher eukaryotes, such as *S. pombe* and *X. laevis*, for which mathematical models are already available. Other simulation tools will soon be available through the web interface to enable more specific analysis such as the automatic bifurcation identification. We also plan to include different modelling approaches, such as Petri nets and Boolean networks, in order to enlarge the simulation possibilities of this resource.

ACKNOWLEDGEMENTS

This work has been supported by the European projects BioinfoGRID, EGEEII, INTAS Ref. Nr 05-1000008-8028 and by MIUR-FIRB Italian projects LITBIO, ITALBIONET, 'Bioinformatics Population Genetics Analysis' and by INGENIO Global Funds delivered by the European Social Fund, by the Ministry of Job, by the Social Welfare and by Regione Lombardia, Italy.

We would like to acknowledge Chiara Bishop for the graphical layout of the website and for proofreading this article, John Hatton for the network management and for the system administration support. Funding to pay the Open Access publication charges for this article was provided by the MIUR-FIRB Italian project ITALBIONET.

Conflict of interest statement. None declared.

REFERENCES

1. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **27**, 29–34.
2. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G.R., Croft,D., De Bono,B., Gillespie,M., Jassal,B. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol*., **8**, R39.
3. Bartlett,R. and Nurse,P. (1990) Yeast as a model system for understanding the control of DNA replication in eukaryotes. *Bioessays*., **12**, 457–463.
4. Hucka,M. *et al*. (2003) The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
5. Le Novere,N., Bornstein,B., Broicher,A., Courtot,M., Donizelli,M., Dharuri,H., Li,L., Sauro,H., Schilstra,M. *et al*. (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res*., **34**, D689–D691.
6. Vass,M.N., Allen,C.A., Shaffer,N., Ramakrishnan,L.T., Watson, and Tyson,J.J. (2004) The JigCell Model Builder and Run Manager. *Bioinformatics*, **20**, 3680–3681.
7. Levene,M. and Loizou,G. (2003) Why is the snowflake schema a good data warehouse design? *Inf. Syst*., **28**, 225–240.
8. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
9. Vermeulen,K., Van Bockstaele,D.R. and Berneman,Z.N. (2003) The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif*., **36**, 131–149.
10. Milanesi,L., Petrillo,M., Sepe,L., Boccia,A., D'Agostino,N., Passamano,M., Di Nardo,S., Tasco,G., Casadio,R. *et al*. (2005) Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity. *BMC Bioinformatics*, **6** (Suppl. 4), S20.
11. Sanner,M.F., Olson,A.J. and Spehner,J.C. (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320.
12. *Mathematical Markup Language (MathML), Version 2.0 (second edition)*.
13. Ermentrout,B. (2002) *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students* SIAM, Philadelphia, USA.
14. Williams,T. & Kelley,C. (1998) GNUPLOT: An Interactive Plotting Program, Version 3.7 organized by: David Denholm. (http://www.gnuplot.info/)
15. Bai,S., Goodrich,D., Thron,C.D., Tecarro,E. and Obeyesekere,M. (2003) Theoretical and experimental evidence for hysteresis in cell proliferation. *Cell Cycle*, **2**, 46–52.